Aligning words using matrix factorisation

Cyril Goutte, Kenji Yamada and Eric Gaussier
Xerox Research Centre Europe
6, chemin de Maupertuis
F-38240 Meylan, France
Cyril.Goutte,Kenji.Yamada,Eric.Gaussier@xrce.xerox.com

Abstract

Aligning words from sentences which are mutual translations is an important problem in different settings, such as bilingual terminology extraction, Machine Translation, or projection of linguistic features. Here, we view word alignment as matrix factorisation. In order to produce proper alignments, we show that factors must satisfy a number of constraints such as orthogonality. We then propose an algorithm for orthogonal non-negative matrix factorisation, based on a probabilistic model of the alignment data, and apply it to word alignment. This is illustrated on a French-English alignment task from the Hansard.

1 Introduction

Aligning words from mutually translated sentences in two different languages is an important and difficult problem. It is important because a word-aligned corpus is typically used as a first step in order to identify phrases or templates in phrase-based Machine Translation (Och et al., 1999), (Tillmann and Xia, 2003), (Koehn et al., 2003, sec. 3), or for projecting linguistic annotation across languages (Yarowsky et al., 2001). Obtaining a word-aligned corpus usually involves training a word-based translation models (Brown et al., 1993) in each directions and combining the resulting alignments.

Besides processing time, important issues are completeness and propriety of the resulting alignment, and the ability to reliably identify general N-to-M alignments. In the following section, we introduce the problem of aligning words from a corpus that is already aligned at the sentence level. We show how this problem may be phrased in terms of matrix factorisation. We then identify a number of constraints on word alignment, show that these constraints entail that word alignment is equivalent to orthogonal non-negative matrix factorisation, and we give a novel algorithm that solves this problem. This is illustrated using data from the shared tasks of the 2003 HLT-NAACL Workshop on Building and Using Parallel Texts (Mihalcea and Pedersen, 2003).

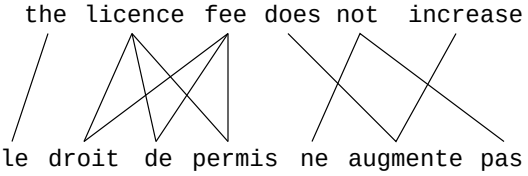

Figure 1: 1-1, M-1, 1-N and M-N alignments.

2 Word alignments

We address the following problem: Given a source sentence $\mathbf{f} = f_1 \dots f_i \dots f_I$ and a target sentence $\mathbf{e} = e_1 \dots e_j \dots e_J$, we wish to find words f_i and e_j on either side which are *aligned*, ie in mutual correspondence. Note that words may be aligned without being directly "dictionary translations". In order to have proper alignments, we want to enforce the following constraints:

Coverage: Every word on either side must be aligned to at least one word on the other side (Possibly taking "null" words into account).

Transitive closure: If f_i is aligned to e_j and e_ℓ, any f_k aligned to e_ℓ must also de aligned to e_j.

Under these constraints, there are only 4 types of alignments: 1-1, 1-N, M-1 and M-N (fig. 1). Although the first three are particular cases where N=1 and/or M=1, the distinction is relevant, because most word-based translation models (eg IBM models (Brown et al., 1993)) can typically not accommodate general M-N alignments.

We formalise this using the notion of *cepts*: a cept is a central pivot through which a subset of e-words is aligned to a subset of f-words. General M-N alignments then correspond to M-1-N alignments from e-words, to a cept, to f-words (fig. 2). Cepts naturally guarantee transitive closure as long as each word is connected to a single cept. In addition, coverage is ensured by imposing that each

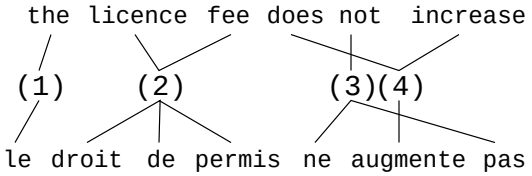

Figure 2: Same as figure 1, using cepts (1)-(4).

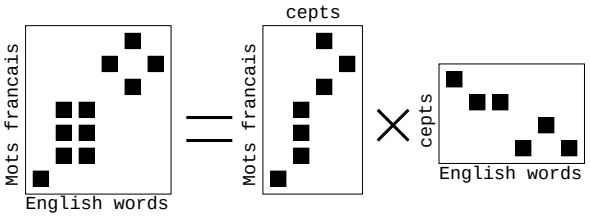

Figure 3: Matrix factorisation of the example from fig. 1, 2. Black squares represent alignments.

word is connected to a cept. A unique constraint therefore guarantees proper alignments:

Propriety: Each word is associated to exactly one cept, and each cept is associated to at least one word on each side.

Note that our use of *cepts* differs slightly from that of (Brown et al., 1993, sec.3), inasmuch cepts may not overlap, according to our definition.

The motivation for our work is that better word alignments will lead to better translation models. For example, we may extract better chunks for phrase-based translation models. In addition, proper alignments ensure that cept-based phrases will cover the entire source and target sentences.

3 Matrix factorisation

Alignments between source and target words may be represented by a $I \times J$ *alignment matrix* $A = [a_{ij}]$, such that $a_{ij} > 0$ if f_i is aligned with e_j and $a_{ij} = 0$ otherwise. Similarly, given K cepts, words to cepts alignments may be represented by a $I \times K$ matrix F and a $J \times K$ matrix E, with positive elements indicating alignments. It is easy to see that matrix $A = F \times E^\top$ then represents the resulting word-to-word alignment (fig. 3).

Let us now assume that we start from a $I \times J$ matrix $\mathcal{M} = [m_{ij}]$, which we call the *translation matrix*, such that $m_{ij} \geq 0$ measures the strength of the association between f_i and e_j (large values mean close association). This may be estimated using a translation table, a count (eg from a N-best list), etc. Finding a suitable alignment matrix A corresponds to finding factors F and E such that:

$$\mathcal{M} \approx F \times S \times E^\top \qquad (1)$$

where without loss of generality, we introduce a diagonal $K \times K$ scaling matrix S which may give different weights to the different cepts. As factors F and E contain only positive elements, this is an instance of *non-negative matrix factorisation*, aka NMF (Lee and Seung, 1999). Because NMF decomposes a matrix into additive, positive components, it naturally yields a sparse representation.

In addition, the propriety constraint imposes that words are aligned to exactly one cept, ie each row of E and F has exactly one non-zero component, a property we call *extreme sparsity*. With the notation $F = [F_{ik}]$, this means that:

$$\forall i, \forall k \neq l, \quad F_{ik}.F_{il} = 0$$

As line i contains a single non-zero element, either F_{ik} or F_{il} must be 0. An immediate consequence is that $\sum_i F_{ik}.F_{il} = 0$: columns of F are *orthogonal*, that is F is an orthogonal matrix (and similarly, E is orthogonal). Finding the best alignment starting from $\mathcal{M}$ therefore reduces to performing a decomposition into *orthogonal non-negative factors*.

4 An algorithm for Orthogonal Non-negative Matrix Factorisation

Standard NMF algorithms (Lee and Seung, 2001) do not impose orthogonality between factors. We propose to perform the Orthogonal Non-negative Matrix Factorisation (ONMF) in two stages: We first factorise $\mathcal{M}$ using Probabilistic Latent Semantic Analysis, aka PLSA (Hofmann, 1999), then we orthogonalise factors using a Maximum A Posteriori (MAP) assignment of words to cepts. PLSA models a joint probability $P(f, e)$ as a mixture of conditionally independent multinomial distributions:

$$P(f, e) = \sum_c P(c)P(f|c)P(e|c) \qquad (2)$$

With $F = [P(f|c)]$, $E = [P(e|c)]$ and $S = \text{diag}(P(c))$, this is exactly eq. 1. Note also that despite the additional matrix S, if we set $E = [P(e, c)]$, then $P(f|e)$ would factor as $F \times E^\top$, the original NMF formulation). All factors in eq. 2 are (conditional) probabilities, and therefore positive, so PLSA also implements NMF.

The parameters are learned from a matrix $\mathcal{M} = [m_{ij}]$ of (f_i, e_j) counts, by maximising the likelihood using the iterative re-estimation formula of the Expectation-Maximisation algorithm (Dempster et al., 1977), cf. fig. 4. Convergence is guaranteed, leading to a non-negative factorisation of $\mathcal{M}$. The second step of our algorithm is to orthogonalise

$$\text{E-step:} \quad P(c|f_i, e_j) = \frac{P(c)P(f_i|c)P(e_j|c)}{\sum_c P(c)P(f_i|c)P(e_j|c)} \quad (3)$$

$$\text{M-step:} \quad P(c) = \frac{1}{N}\sum_{ij} m_{ij} P(c|f_i, e_j) \quad (4)$$

$$\text{M-step:} \quad P(f_i|c) \propto \sum_j m_{ij} P(c|f_i, e_j) \quad (5)$$

$$\text{M-step:} \quad P(e_j|c) \propto \sum_i m_{ij} P(c|f_i, e_j) \quad (6)$$

Figure 4: The EM algorithm iterates these E and M-steps until convergence.

the resulting factors. Each source word f_i is assigned the most probable cept, ie cept c for which $P(c|f_i) \propto P(c)P(f_i|c)$ is maximal. Factor F is therefore set to:

$$F_{ik} \propto \begin{cases} 1 & \text{if } k = \text{argmax}_c\, P(c|f_i) \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where proportionality ensures that column of F sum to 1, so that F stays a conditional probability matrix. We proceed similarly for target words e_j to orthogonalise E. Thanks to the MAP assignment, each line of F and E contains exactly one non-zero element. We saw earlier that this is equivalent to having orthogonal factors. The result is therefore an orthogonal, non-negative factorisation of the original translation matrix $\mathcal{M}$.

4.1 Number of cepts

In general, the number of cepts is unknown and must be estimated. This corresponds to choosing the number of components in PLSA, a classical model selection problem. The likelihood may not be used as it always increases as components are added. A standard approach to optimise the complexity of a mixture model is to maximise the likelihood, penalised by a term that increases with model complexity, such as AIC (Akaike, 1974) or BIC (Schwartz, 1978). BIC asymptotically chooses the correct model size (for complete models), while AIC always overestimates the number of components, but usually yields good predictive performance. As the largest possible number of cepts is $\min(I, J)$, and the smallest is 1 (all f_i aligned to all e_j), we estimate the optimal number of cepts by maximising AIC or BIC between these two extremes.

4.2 Dealing with null alignments

Alignment to a "null" word may be a feature of the underlying statistical model (eg IBM models), or it may be introduced to accommodate words which have a low association measure with all other words. Using PLSA, we can deal with null alignments in a principled way by introducing a null word on each side (f_0 and e_0), and two null cepts ("f-null" and "e-null") with a 1-1 alignment to the corresponding null word, to ensure that null alignments will only be 1-N or M-1. This constraint is easily implemented using proper initial conditions in EM. Denoting the null cepts as $c_{f\emptyset}$ and $c_{e\emptyset}$, 1-1 alignments between null cepts and the corresponding null words impose the conditions:

1. $P(f_0|c_{f\emptyset}) = 1$ and $\forall i \neq 0, P(f_i|c_{f\emptyset}) = 0$;
2. $P(e_0|c_{e\emptyset}) = 1$ and $\forall j \neq 0, P(e_j|c_{e\emptyset}) = 0$.

Stepping through the E-step and M-step equations (3–6), we see that these conditions are preserved by each EM iteration. In order to deal with null alignments, the model is therefore augmented with two null cepts, for which the probabilities are initialised according to the above conditions. As these are preserved through EM, we maintain proper 1-N and M-1 alignments to the null words. The main difference between null cepts and the other cepts is that we relax the propriety constraint and do not force null cepts to be aligned to at least one word on either side. This is because in many cases, all words from a sentence can be aligned to non-null words, and do not require any null alignments.

4.3 Modelling noise

Most elements of $\mathcal{M}$ usually have a non-zero association measure. This means that for proper alignments, which give zero probability to alignments outside identified blocks, actual observations have exactly 0 probability, ie the log-likelihood of parameters corresponding to proper alignments is undefined. We therefore refine the model, adding a noise component indexed by $c = 0$:

$$\begin{aligned} P(f, e) &= \sum_{c>0} P(c)P(f|c)P(e|c) \\ &\quad + P(c=0)P(f, e|c=0) \end{aligned}$$

The simplest choice for the noise component is a uniform distribution, $P(f, e|c = 0) \propto 1$. E-step and M-steps in eqs. (3–6) are unchanged for $c > 0$, and the E-step equation for $c = 0$ is easily adapted: $P(c{=}0|f, e) \propto P(c{=}0)P(f, e|c{=}0)$.

5 Example

We first illustrate the factorisation process on a simple example. We use the data provided for

the French-English shared task of the 2003 HLT-NAACL Workshop on Building and Using Parallel Texts (Mihalcea and Pedersen, 2003). The data is from the Canadian Hansard, and reference alignments were originally produced by Franz Och and Hermann Ney (Och and Ney, 2000). Using the entire corpus (20 million words), we trained English→French and French→English IBM4 models using GIZA++. For all sentences from the trial and test set (37 and 447 sentences), we generated up to 100 best alignments for each sentence and in each direction. For each pair of source and target words (f_i, e_j), the association measure m_{ij} is simply the number of times these words were aligned together in the two N-best lists, leading to a count between 0 (never aligned) and 200 (always aligned).

We focus on sentence 1023, from the trial set. Figure 5 shows the reference alignments together with the generated counts. There is a background "noise" count of 3 to 5 (small dots) and the largest counts are around 145-150. The N-best counts seem to give a good idea of the alignments, although clearly there is no chance that our factorisation algorithm will recover the alignment of the two instances of '*de*' to '*need*', as there is no evidence for it in the data. The ambiguity that the factorisation will have to address, and that is not easily resolved using, eg, thresholding, is whether '*ont*' should be aligned to '*They*' or to '*need*'.

The N-best count matrix serves as the *translation matrix*. We estimate PLSA parameters for $K = 1 \ldots 6$, and find out that AIC and BIC reach their maximum for $K = 6$. We therefore select 6 cepts for this sentence, and produce the alignment matrices shown on figure 6. Note that the order of the cepts is arbitrary (here the first cept correspond 'et' — 'and'), except for the null cepts which are fixed. There is a fixed 1-1 correspondence between these null cepts and the corresponding null words on each side, and only the French words '*de*' are mapped to a null cept. Finally, the estimated noise level is $P(c = 0) = 0.00053$. The ambiguity associated with aligning '*ont*' has been resolved through cepts 4 and 5. In our resulting model, $P(c{=}4|$'*ont*'$) \approx 0.40$ while $P(c{=}6|$'*ont*'$) \approx 0.54$: The MAP assignment forces '*ont*' to be aligned to cept 5 only, and therefore to '*need*'.

Note that although the count for (*need*,*ont*) is slightly larger than the count for (*they*,*ont*) (cf. fig. 5), this is not a trivial result. The model was able to resolve the fact that *they* and *need* had to be aligned to 2 different cepts, rather than eg a larger cept corresponding to a 2-4 alignment, and to produce proper alignments through the use of these cepts.

6 Experiments

In order to perform a more systematic evaluation of the use of matrix factorisation for aligning words, we tested this technique on the full trial and test data from the 2003 HLT-NAACL Workshop. Note that the reference data has both "Sure" and "Probable" alignments, with about 77% of all alignments in the latter category. On the other hand, our system proposes only one type of alignment. The evaluation is done using the performance measures described in (Mihalcea and Pedersen, 2003): precision, recall and F-score on the probable and sure alignments, as well as the *Alignment Error Rate* (AER), which in our case is a weighted average of the recall on the sure alignments and the precision on the probable. Given an alignment $\mathcal{A}$ and gold standards $\mathcal{G}_S$ and $\mathcal{G}_P$ (for sure and probable alignments, respectively):

$$P_T = \frac{|\mathcal{A} \cap \mathcal{G}_T|}{|\mathcal{A}|} \quad (8)$$

$$R_T = \frac{|\mathcal{A} \cap \mathcal{G}_T|}{|\mathcal{G}_T|} \quad (9)$$

$$F_T = \frac{2P_T R_T}{P_T + R_T} = \frac{2|\mathcal{A} \cap \mathcal{G}_T|}{|\mathcal{G}_T| + |\mathcal{A}|} \quad (10)$$

where T is either S or P, and:

$$\text{AER} = 1 - \frac{|\mathcal{G}_S|R_S + |\mathcal{A}|P_P}{|\mathcal{G}_S| + |\mathcal{A}|} \quad (11)$$

Using these measures, we first evaluate the performance on the trial set (37 sentences): as we produce only one type of alignment and evaluate against "Sure" and "Probable", we observe, as expected, that the recall is very good on sure alignments, but precision relatively poor, with the reverse situation on the probable alignments (table 1). This is because we generate an intermediate number of alignments. There are 338 sure and 1446 probable alignments (for 721 French and 661 English words) in the reference trial data, and we produce 707 (AIC) or 766 (BIC) alignments with ONMF. Most of them are at least probably correct, as attested by P_P, but only about half of them are in the "Sure" subset, yielding a low value of P_S. Similarly, because "Probable" alignments were generated as the union of alignments produced by two annotators, they sometimes lead to very large M-N alignments, which produce on average 2.5 to 2.7 alignments per word. By contrast ONMF produces less than 1.2 alignments per word, hence the low value of R_P. As the AER is a weighted average of R_S and P_P, the resulting AER are relatively low for our method.

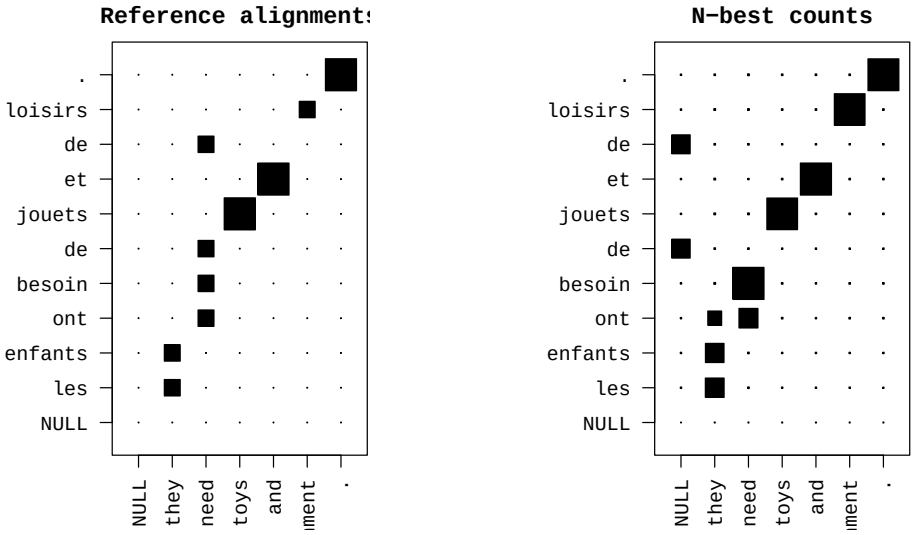

Figure 5: Left: reference alignments, large squares are sure, medium squares are probable; Right: accumulated counts from IBM4 N-best lists, bigger squares are larger counts.

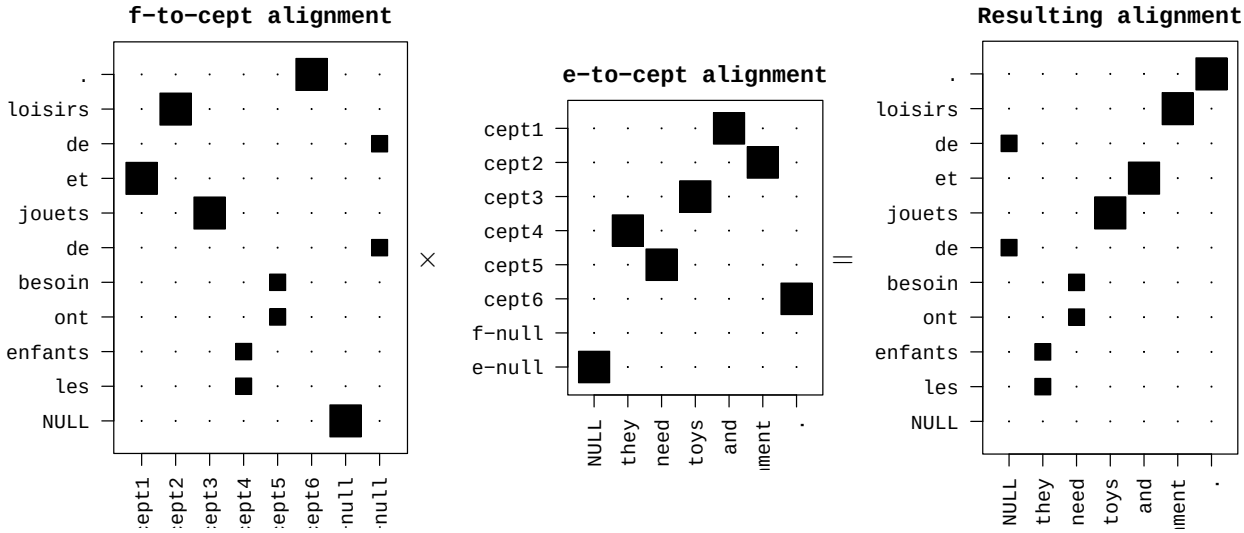

Figure 6: Resulting word-to-cept and word-to-word alignments for sentence 1023.

Method	P_S	R_S	F_S	P_P	R_P	F_P	AER
ONMF + AIC	45.26%	94.67%	61.24%	86.56%	34.30%	49.14%	10.81%
ONMF + BIC	42.69%	96.75%	59.24%	83.42%	35.82%	50.12%	12.50%

Table 1: Performance on the 37 trial sentences for orthogonal non-negative matrix factorisation (ONMF) using the AIC and BIC criterion for choosing the number of cepts, discounting null alignments.

We also compared the performance on the 447 test sentences to 1/ the intersection of the alignments produced by the top IBM4 alignments in either directions, and 2/ the best systems from (Mihalcea and Pedersen, 2003). On limited resources, Ralign.EF.1 (Simard and Langlais, 2003) produced the best F-score, as well as the best AER when NULL alignments were taken into account, while XRCE.Nolem.EF.3 (Dejean et al., 2003) produced the best AER when NULL alignments were discounted. Tables 2 and 3 show that ONMF improves on several of these results. In particular, we get better recall and F-score on the probable alignments (and even a better precision than Ralign.EF.1 in table 2). On the other hand, the performance, and in particular the precision, on sure alignments is dismal. We attribute this at least partly to a key difference between our model and the reference data:

Method	P_S	R_S	F_S	P_P	R_P	F_P	AER
ONMF + AIC	49.86%	95.12%	65.42%	84.63%	37.39%	51.87%	11.76%
ONMF + BIC	46.50%	**96.01%**	62.65%	80.92%	**38.69%**	**52.35%**	14.16%
IBM4 intersection	71.46%	90.04%	**79.68%**	**97.66%**	28.44%	44.12%	**5.71%**
HLT-03 best F	**72.54%**	80.61%	76.36%	77.56%	38.19%	51.18%	18.50%
HLT-03 best AER	55.43%	93.81%	69.68%	90.09%	35.30%	50.72%	8.53%

Table 2: Performance on the 447 English-French test sentences, discounting NULL alignments, for orthogonal non-negative matrix factorisation (ONMF) using the AIC and BIC criterion for choosing the number of cepts. HLT-03 best F is Ralign.EF.1 and best AER is XRCE.Nolem.EF.3 (Mihalcea and Pedersen, 2003).

our model enforces coverage and makes sure that all words are aligned, while the "Sure" reference alignments have no such constraints and actually have a very bad coverage. Indeed, less than half the words in the test set have a "Sure" alignment, which means that a method which ensures that all words are aligned will at best have a sub 50% precision. In addition, many reference "Probable" alignments are not proper alignments in the sense defined above.

Note that the IBM4 intersection has a bias similar to the sure reference alignments, and performs very well in F_S, P_P and especially in AER, even though it produces very incomplete alignments. This points to a particular problem with the AER in the context of our study. In fact, a system that outputs exactly the set of sure alignments achieves a perfect AER of 0, even though it aligns only about 23% of words, clearly an unacceptable drawback in many applications. We think that this issue may be addressed in two different ways. One time-consuming possibility would be to post-edit the reference alignment to ensure coverage and proper alignments. Another possibility would be to use the probabilistic model to mimic the reference data and generate both "Sure" and "Probable" alignments using eg thresholds on the estimated alignment probabilities. This approach may lead to better performance according to our metrics, but it is not obvious that the produced alignments will be more reasonable or even useful in a practical application.

We also tested our approach on the Romanian-English task of the same workshop, cf. table 4. The 'HLT-03 best' is our earlier work (Dejean et al., 2003), simply based on IBM4 alignment using an additional lexicon extracted from the corpus. Slightly better results have been published since (Barbu, 2004), using additional linguistic processing, but those were not presented at the workshop. Note that the reference alignments for Romanian-English contain only "Sure" alignments, and therefore we only report the performance on those. In addition, AER $= 1 - F_S$ in this setting. Table 4 shows that the matrix factorisation approach does not offer any quantitative improvements over these results. A gain of up to 10 points in recall does not offset a large decrease in precision. As a consequence, the AER for ONMF+AIC is about 10% higher than in our earlier work. This seems mainly due to the fact that the 'HLT-03 best' produces alignments for only about 80% of the words, while our technique ensure coverage and therefore aligns *all* words. These results suggest that remaining 20% seem particularly problematic. These quantitative results are disappointing given the sofistication of the method. It should be noted, however, that ONMF provides the qualitative advantage of producing proper alignments, and in particular ensures coverage. This may be useful in some contexts, eg training a phrase-based translation system.

7 Discussion

7.1 Model selection and stability

Like all mixture models, PLSA is subject to local minima. Although using a few random restarts seems to yield good performance, the results on difficult-to-align sentences may still be sensitive to initial conditions. A standard technique to stabilise the EM solution is to use deterministic annealing or tempered EM (Rose et al., 1990). As a side effect, deterministic annealing actually makes model selection easier. At low temperature, all components are identical, and they differentiate as the temperature increases, until the final temperature, where we recover the standard EM algorithm. By keeping track of the component differentiations, we may consider multiple effective numbers of components in one pass, therefore alleviating the need for costly multiple EM runs with different cept numbers and multiple restarts.

7.2 Other association measures

ONMF is only a tool to factor the original translation matrix $\mathcal{M}$, containing measures of associations between f_i and e_j. The quality of the resulting alignment greatly depends on the way $\mathcal{M}$ is

Method	P_S	R_S	F_S	P_P	R_P	F_P	AER
ONMF + AIC	42.88%	95.12%	59.11%	75.17%	37.20%	49.77%	18.63%
ONMF + BIC	40.17%	**96.01%**	56.65%	72.20%	**38.49%**	**50.21%**	20.78%
IBM4 intersection	56.39%	90.04%	69.35%	**81.14%**	28.90%	42.62%	**15.43%**
HLT-03 best	**72.54%**	80.61%	**76.36%**	77.56%	36.79%	49.91%	18.50%

Table 3: Performance on the 447 English-French test sentences, taking NULL alignments into account, for orthogonal non-negative matrix factorisation (ONMF) using the AIC and BIC criterion for choosing the number of cepts. HLT-03 best is Ralign.EF.1 (Mihalcea and Pedersen, 2003).

	no NULL alignments				with NULL alignments			
Method	P_S	R_S	F_S	AER	P_S	R_S	F_S	AER
ONMF + AIC	70.34%	65.54%	67.85%	32.15%	62.65%	62.10%	62.38%	37.62%
ONMF + BIC	55.88%	**67.70%**	61.23%	38.77%	51.78%	**64.07%**	57.27%	42.73%
HLT-03 best	**82.65%**	62.44%	**71.14%**	**28.86%**	**82.65%**	54.11%	**65.40%**	**34.60%**

Table 4: Performance on the 248 Romanian-English test sentences (only sure alignments), for orthogonal non-negative matrix factorisation (ONMF) using the AIC and BIC criterion for choosing the number of cepts. HLT-03 best is XRCE.Nolem (Mihalcea and Pedersen, 2003).

filled. In our experiments we used counts from N-best alignments obtained from IBM model 4. This is mainly used as a proof of concept: other strategies, such as weighting the alignments according to their probability or rank in the N-best list would be natural extensions. In addition, we are currently investigating the use of translation and distortion tables obtained from IBM model 2 to estimate $\mathcal{M}$ at a lower cost. Ultimately, it would be interesting to obtain association measures m_{ij} in a fully non-parametric way, using corpus statistics rather than translation models, which themselves perform some kind of alignment. We have investigated the use of co-occurrence counts or mutual information between words, but this has so far not proved successful, mostly because common words, such as function words, tend to dominate these measures.

7.3 M-1-0 alignments

In our model, cepts ensure that resulting alignments are proper. There is however one situation in which improper alignments may be produced: If the MAP assigns f-words but no e-words to a cept (because e-words have more probable cepts), we may produce "orphan" cepts, which are aligned to words only on one side. One way to deal with this situation is simply to remove cepts which display this behaviour. Orphaned words may then be re-assigned to the remaining cepts, either directly or after re-training PLSA on the remaining cepts (this is guaranteed to converge as there is an obvious solution for $K = 1$).

7.4 Independence between sentences

One natural comment on our factorisation scheme is that cepts should not be independent between sentences. However it is easy to show that the *factorisation* is optimally done on a sentence per sentence basis. Indeed, what we factorise is the association measures m_{ij}. For a sentence-aligned corpus, the association measure between source and target words from two different sentence pairs should be exactly 0 because words should not be aligned across sentences. Therefore, the larger translation matrix (calculated on the entire corpus) is block diagonal, with non-zero association measures only in blocks corresponding to aligned sentence. As blocks on the diagonal are mutually orthogonal, the optimal global orthogonal factorisation is identical to the block-based (ie sentence-based) factorisation. Any corpus-induced dependency between alignments from different sentences must therefore be built in the association measure m_{ij}, and cannot be handled by the factorisation method. Note that this is the case in our experiments, as model 4 alignments rely on parameters obtained on the entire corpus.

8 Conclusion

In this paper, we view word alignment as 1/ estimating the association between source and target words, and 2/ factorising the resulting association measure into orthogonal, non-negative factors. For solving the latter problem, we propose an algorithm for ONMF, which guarantees both proper alignments and good coverage. Experiments carried out on the Hansard give encouraging results, in the sense that

we improve in several ways over state-of-the-art results, despite a clear bias in the reference alignments. Further investigations are required to apply this technique on different association measures, and to measure the influence that ONMF may have, eg on a phrase-based Machine Translation system.

Acknowledgements

We acknowledge the Machine Learning group at XRCE for discussions related to the topic of word alignment. We would like to thank the three anonymous reviewers for their comments.

References

H. Akaike. 1974. A new look at the statistical model identification. *IEEE Tr. Automatic Control*, 19(6):716–723.

A.-M. Barbu. 2004. Simple linguistic methods for improving a word alignment algorithm. In *Le poids des mots — Proc. JADT04*, pages 88–98.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19:263–312.

H. Dejean, E. Gaussier, C. Goutte, and K. Yamada. 2003. Reducing parameter space for word alignment. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts*, pages 23–26.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society, Series B*, 39(1):1–38.

T. Hofmann. 1999. Probabilistic latent semantic analysis. In *Uncertainty in Artificial Intelligence*, pages 289–296.

P. Koehn, F. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. HLT-NAACL 2003*.

D. D. Lee and H. S. Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791.

D. D. Lee and H. S. Seung. 2001. Algorithms for non-negative matrix factorization. In *NIPS*13*, pages 556–562.

R. Mihalcea and T. Pedersen. 2003. An evaluation exercise for word alignment. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts*, pages 1–10.

F. Och and H. Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proc. COLING'00*, pages 1086–1090.

F. Och, C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine translation. In *Proc. EMNLP*, pages 20–28.

K. Rose, E. Gurewitz, and G. Fox. 1990. A deterministic annealing approach to clustering. *Pattern Recognition Letters*, 11(11):589–594.

G. Schwartz. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

M. Simard and P. Langlais. 2003. Statistical translation alignment with compositionality constraints. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts*, pages 19–22.

C. Tillmann and F. Xia. 2003. A phrase-based unigram model for statistical machine translation. In *Proc. HLT-NAACL 2003*.

D. Yarowsky, G. Ngai, and R. Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proc. HLT 2001*.